AUTOMATIC ANALYSIS AND RECOGNITION OF INTONATION PATTERNS - A FEASIBILITY STUDY

A Thesis Submitted
In Partial Fulfilment of the Requirements
for the Degree of
MASTER OF TECHNOLOGY

By

S. K. SONI

to the

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY KANPUR

SEPTEMBER, 1973

EE-1973-M-SON-AUT

Thesis
621.381958
So58

I. I. T. KANPUR
CENTRAL LIBRARY
Acc. No. A 26282

26 SEP 1973

V
JUNE '76

POST GRADUATE OFFICE
Submitted on 7.9.73

CERTIFICATE

This is to certify that the thesis entitled, "Automatic Analysis and Recognition of Intonation Patterns - A Feasibility Study" is a record of the work carried out under my supervision and that this has not been submitted elsewhere for a degree.

Kanpur
September 1973

N. Ramasubramanian
Assistant Professor
Electrical Engineering Department
Indian Institute of Technology, Kanpur

POST GRADUATE OFFICE
This thesis has been approved
for the award of the Degree of
Master of Technology (M.Tech.)
in accordance with the
regulations of the Indian
Institute of Technology Kanpur
Dated. 17.9.73

ACKNOWLEDGEMENT

I sincerely thank Dr. N. Ramasubramaniam for having suggested this interesting problem and for his guidance and encouragement during the work.

I wish to express profound sense of gratitude to Dr. V.P. Sinha for his consistent help throughout the work. But for his assistance my efforts would have been futile. I thank him for permitting to make use of some of his computer programs on digital filtering of time series.

My acknowledgements are due to the staff of Electrical Engineering Department and Computer Centre for the co-operation in carrying out this work successfully.

S. K. Soni

ABSTRACT

The intonation of a speech is an indication of the type of statement of the sentence eg. an interrogative, a declarative etc. Many important inference can be deduced by knowing the intonation pattern in a continuous speech recognition system. Intonation are of two types: word intonation and sentence intonation. A method has been developed to find out the sentence intonation here.

Vibrations of vocal cords result in the fundamental frequency (f_o) of speech. The variation of f_o with time is the intonation contour.

The variation of f_o has been evaluated with respect to time by digital computation. The analog speech signal after conversion to digital form is processed through a bank of band pass filters simulated in the digital computer. The output of this bank is stored in a C-matrix. The contents of C-matrix indicate the intonation pattern of the utterance.

TABLE OF CONTENTS

CHAPTER	TITLE	Page
1	INTRODUCTION	1
2	REVIEW OF LITERATURE	11
3	STATEMENT OF PROBLEM AND PROPOSED SOLUTION	17
4	SIMULATED ANALYZER	21
5	RESULTS OF EXPERIMENTS	31
6	CONCLUSION	36
7	APPENDIX A	41
8	APPENDIX B	49
9	BIBLIOGRAPHY	51

CHAPTER 1

INTRODUCTION

1.1 Motivation

With the vast development in the digital computers, the communications with machines have already made the transition from science fiction to a practical reality. Modern computers now can handle problems considered to be non-trivial and complex by human beings. This has opened the path for more versatile, 'natural' man-computer communication techniques. Thus recent years have seen the progress of time sharing systems for rapid man-machine interaction; graphical display of pictures, graphs, languages on cathode ray tubes; and is discussed herein an approach for man-machine communication by speech(voice). Among the various channels available for communication with machine, the speech has an edge over the others [23].

A few advantages of speech channel are: it provides most effortless encoding of all output channels, has higher data rate compared to other output channels, prefarable for spontaneous output, does not tie up hands, eyes, feet or ears; can be used while in motion, requires inexpensive and readily available terminal equipment.

Since the future man-machine communication would be desirable through voice, due to the above advantages

and since a computer is a most versatile machine we are interested here to findout how digital computers can be used in speech recognition systems.

1.2 Role of Computer in Speech Recognition

The speech recognition by computers can be viewed from two angles.

(a) The contents of the message

(b) The linguistic aspects of the message.

In the former, the words used in the speech are analysed to gather certain information for the purposes of understanding them (called the semantics). For example, some particular words (say, digits) are used to perform certain functions [28]. These words spoken by a person could be stored suitably in a table or array in a computer memory and later on, when spoken again, they could be compared with the elements of the table and if matched, certain chain of actions can be initiated. The linguistic aspects, like, the stress on vowel, tone and intonation are not utilized herein. However, it forms part of type (b). Though much work has been done in this aspect of speech recognition [1,4,9,21] yet because of the 'uncertain nature' (or dependency on number of factors like emotional state, sex, age etc.) of the speech, much more is required to enhance the man-machine communication. Of immediate concern to us here is the intonation. We would venture to extract the information about the intonation of speech

signal which can then be classified to indicate various types of statements.

1.3 Definition: Intonation

Intonation may be defined as the gross variation of the voice fundamental frequency over a sentence or phrase where a phrase or sentence is made up of one or more word strings in a given language.

Both the phrase and sentence intonation carry important information from speech recognition point of view. What we shall be considering in the present work is the sentence intonation. As will be seen shortly the various intonation patterns would be obtained for different types of statements and hence their classification.

1.4 Speech Production Mechanism

The generation of intonation of an utterance is organized in parts, in terms of certain synchronized patterns of muscular activities of the larynx and the respiratory system. Intonation is therefore perceived in terms of complete contour of fundamental frequency and amplitude. The sounds are produced by a few basic mechanisms obtained in various ways. A regular periodic vibration is generated through the action of the vocal cords, which are two bands of elastic tissues in the larynx. They may be opened to permit free breathing or brought together to produce various types of voice qualities. At a given moment, they are closed tightly

with narrow opening between them and if the air behind is moderately compressed by the contraction of thoraic cavity, the air pressure forces the vocal cords apart. A small amount of air passes through them and they again are closed because of their elastic nature. And thus the alternate closing and opening of the vocal cords give rise to voice fundamental frequency(f_o).

Based on the above facts the concept of source-filter model of speech production [7] was developed, as shown in Figure 1.

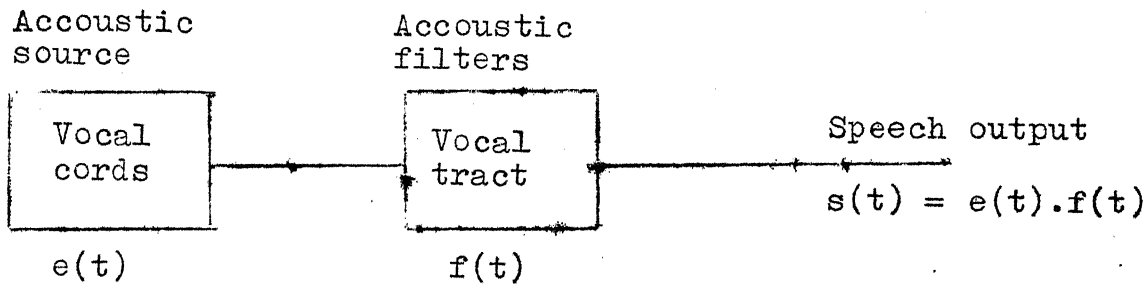


Figure 1.1: A simplified model of speech production system.

The vocal tract is excited by vibrations of the vocal cords modulating the air stream passing through the larynx. The variation in volume-velocity of the air stream is the quasi-periodic pulse like waveform that is ultimately responsible for the fundamental frequency of the voice. The vocal tract consists of oral, nasal and pharyngeal cavities. It may be considered as a linear filter network with a fundamental frequency waveform as its input and the radiated speech signal as its output [8].

From the communication point of view [20] physiologically in general, for males the fundamental frequency varies from 70 Hz to 160 Hz; for females 140 Hz to 320 Hz, and upto 500 Hz for children. Health factors affecting the f_o here would be assumed to be inconsequential. The effect of emotions on the fundamental frequency f_o have been studied extensively by others [20].

Sentence intonation is nothing but the **contour** of the fundamental frequency f_o. Therefore measurement of f_o over the period of utterance would result in the formation of intonatuon patterns.

1.5 Types of Intonation Patterns

The various types of sentences may be distinguished from the study of the variation of f_o during the utterance. Terminal glide in the f_o gives an important clue to the type of the sentence. Questions are often distinguished from statements by a terminal rise in f_o as against a terminal fall in a statement. However, a question may also be distinguished by a comparatively high f_o throughout the utterance [16]. Spectrographic analysis of speech also have shown that, questions tend to be spoken on a higher f_o than statements usually ending in a moderate rise [16]. These facts suggest that not only the direction and range of terminal glide (rise and fall) but the shape and level of the entire contour affect the listner's judgement as to the type of sentences perceived. Broadly

speaking the sentences could be divided into the following major categories:

1. Normal statement
2. Questions
3. Anger
4. Delighted surprise (exclamation).

As shown elsewhere [12], the average f_o patterns for the intonation of above 4 types could be illustrated as in Figure 1.2. The utterance chosen was 'five thousand six hundred ten' spoken in different ways. The variation of f_o at the terminal glide is marked in the case of question

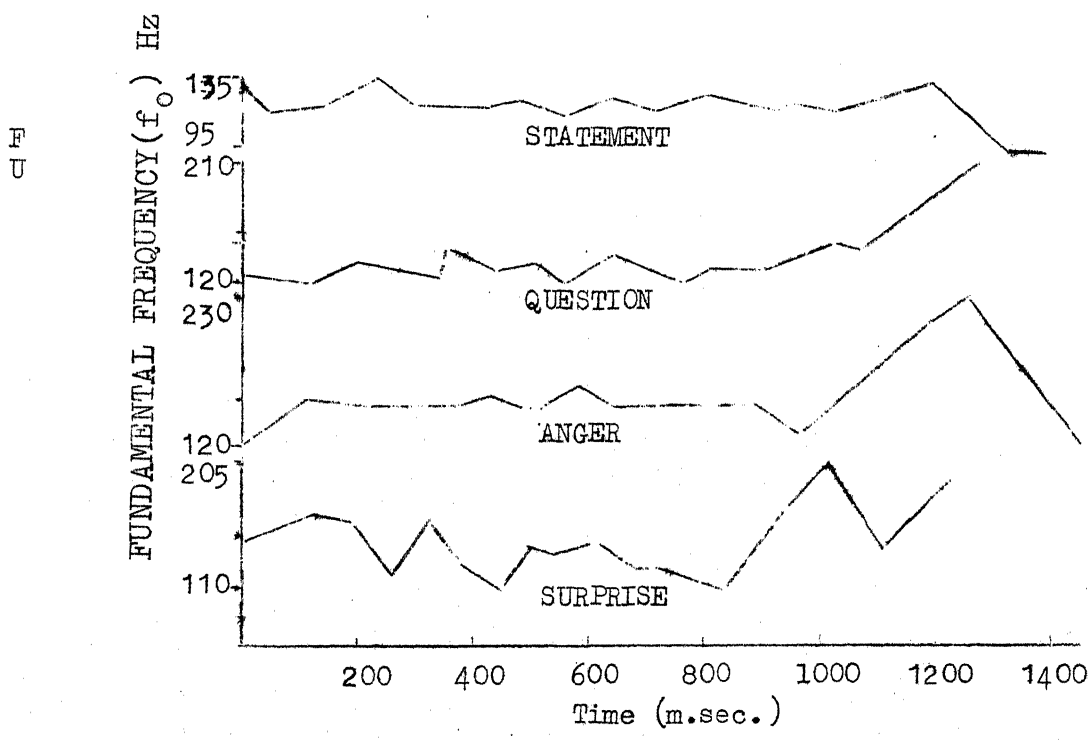


Figure 1.2: Average fundamental frequency patterns of the intonation contours.

Based on these patterns we can analyse the frequency component of an utterance with respect to time and depending upon the type we may classify the type of sentences.

1.6 A Recognition System

It will be appropriate here to see one example of Man-computer communication system through voice and orient our problem of intonation in this environment.

Let us consider a block diagram of a general purpose man-computer communication system through voice [26] as in Figure 1.3.

It is clear from the block diagram that the input speech after digitization is stored in the computer memory. This is then analysed through suitable spectrum analyzer. The output of the analyzed speech is recognized as a string of sounds (phonemes) and then segmented to form sentences and suitable classification is done thereby in phase V; the message reconition interpreation of the speech is done and appropriate responses are generated suitably and outputed as desired bythe interactor.Careful study of literature on speech analysis [8] shows that even with different types of spectrum analyzers often it is difficult to resolve the ambiguities of recognition of a different classes of speech sounds, e.g. a vowel /O/ and a glide /w/ might be confused on the basis of their gross spectral information - formants, duration etc. To resolve these

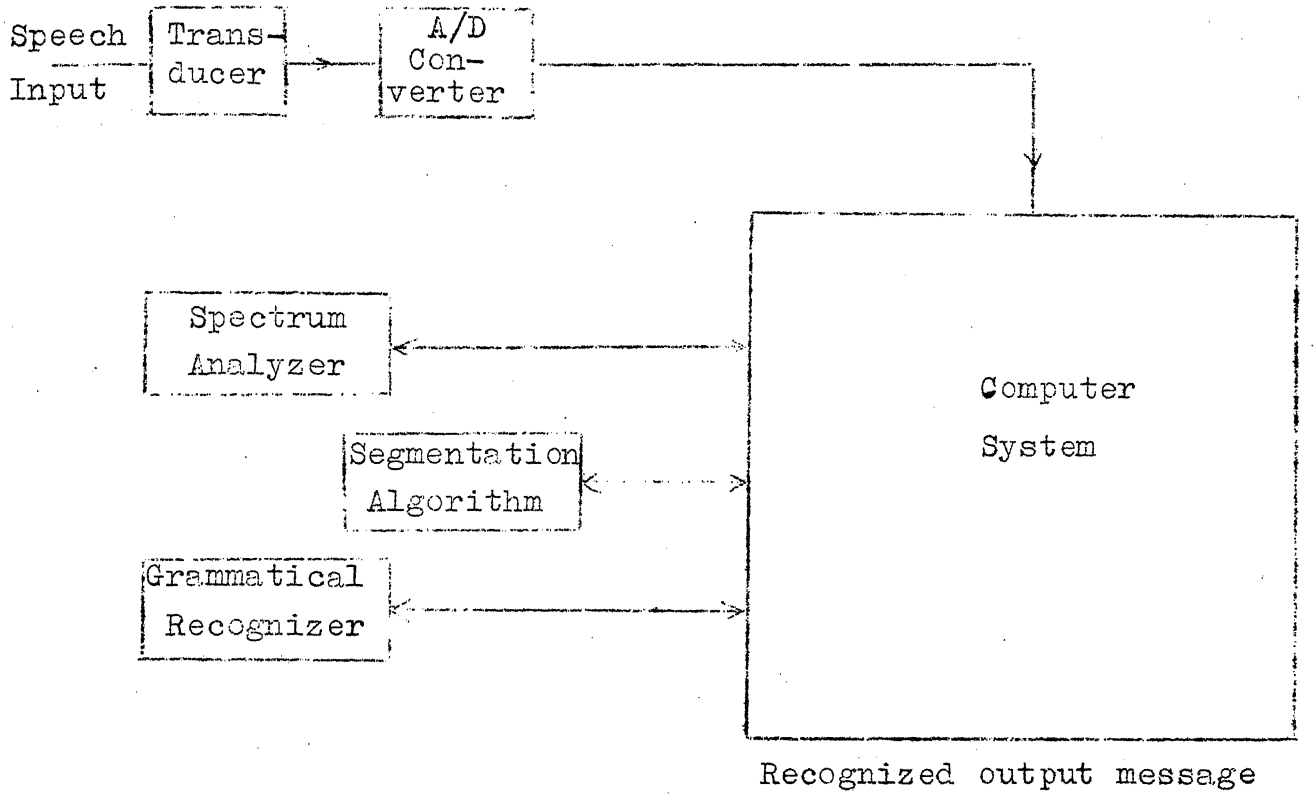


Figure 1.3: Block diagram of a speech recognition system.

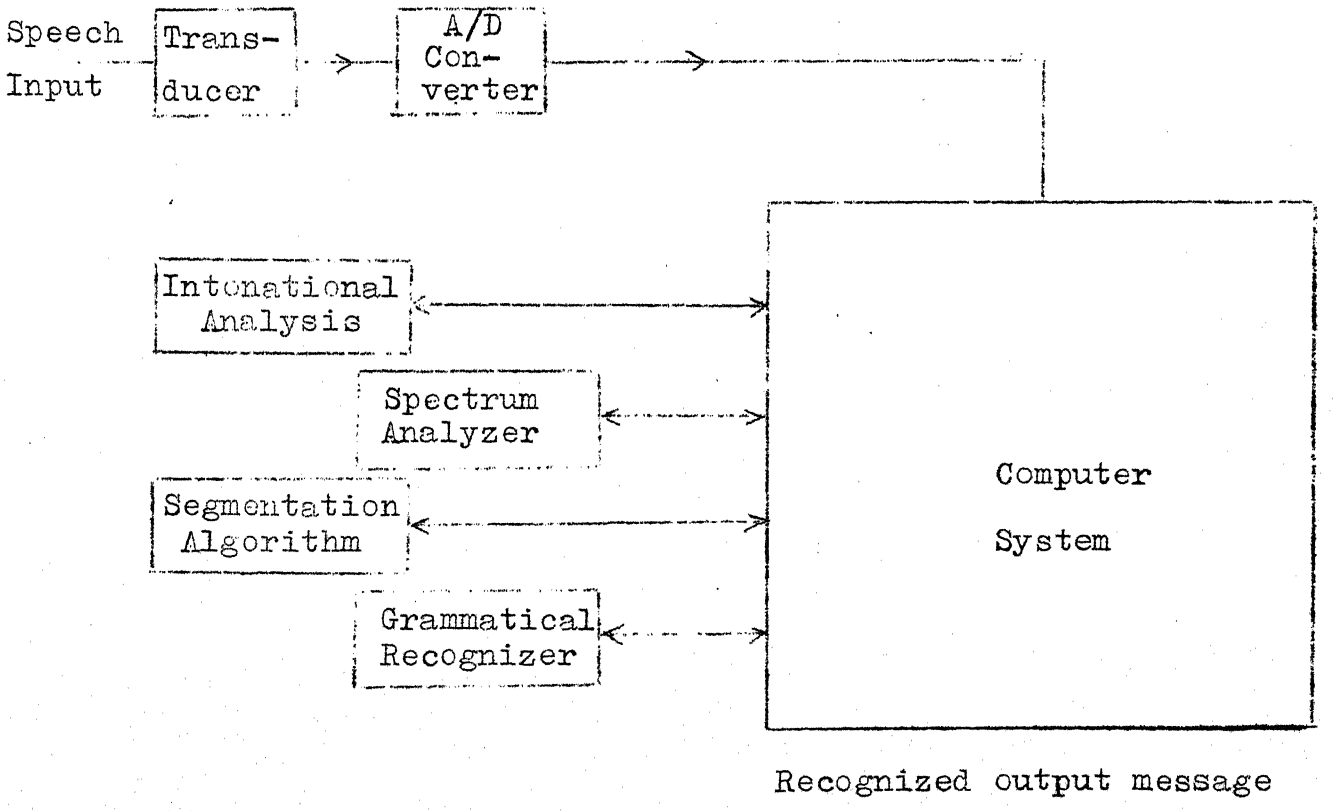


Figure 1.4: Modified speech recognition system using intonational analysis.

difficulties, one may think of two of the following approaches:

(1) ask the speaker to repeatedly speak the message very slowly and clearly. However, this may not always be possible,

(2) take the cues from other acoustic-linguistic aspects- say, intonation contours.

The second seems to be viable because even though the recognition of a particular speech sound may be difficult in a given sample, since an intonation type is always present in a speech, this can be utilized suitably. To illustrate this point, let us take an example:

Suppose a normal man with normal organs of speech (vocal set up) speaks a sentence in English: say (we have transcribed in normal orthography here).

When did you come ?

If on analysis, we note that the speech segment /w/ could not be clearly, and unambiguously (say undifferentiable from /O/ or /v/) identified, what one can do is as follows:

1. Look at the intonation contour.
2. Does it raise at the end of the utterance ? if yes, and if the second segment is identified as /h/, and a vowel after it, then the utterance begins with wh - (who, when etc.).
3. Otherwise call other recognition algorithms.

Without elaborating further, we note that in English, normally a querry sentence ends with raising terminal

intonation, irrespective of whether a querry word such as what, when, which, who, occurs at the beginning or not. Thus one can construct a complex algorithm suitably incorporating these facts. The important point to be noted here is, that the intonation analysis of a speech input could be really used to call appropriate types of spectrum analysis techniques (hardware or software wise) and thus enhance the recognition procedure enormously. Secondly, while the normal speech analysis techniques restrict their attention to the recognition of vowels, syllables/words, this approach would open procedures for continuous recognition of speech. To our knowledge, we are first to introduce this approach in speech recognition system.

In Figure 1.4, we have provided the modified version of Figure 1.3 system incorporating the intonation analysis procedure before the spectrum analyzer is used.

In this thesis an attempt has been made to establish the feasibility of the analysis, recognition and classification of intonation patterns as a first step in the direction of total continuous speech recognition. Using a bank of simulated digital filters for this purpose, a preliminary investigation has been done with the synthesized signal samples. In Chapters 3 and 4 digital filter simulation done on IBM 7044 is described and the problems encountered and the suggestions for further work are given in Chapter 6.

CHAPTER 2

REVIEW OF LITERATURE

Speech research studies have gained momentum since 1960. Comprehensive survey of literature on the automatic speech recognition system exists [17]. This can be further supplemented suitably to include materials published upto 1973 (Appendix A). Most of the studies can be broadly classified as follows:

(a) Isolated spoken vowel recognition

(b) Syllable recognition

(c) Numeral (digits) recognition

(d) Word recognition

(e) Speaker recognition

(f) Prosodic feature recognition (stress and tone).

Apart from the areas of speech recognition different techniques, hardware and software have emerged in hthe recent past. A few broad systems can be summarized as follows:

(a) Hardware recognition system

(b) Hardware cum computer combination

(c) Adaptive (learning) recognizers

(d) Usual accoustic analysis

(e) Simulated recognizers.

In the following brief outline of the various recognition systems we shall mention only the most significant ones and rest of them may be found in the literature [17]. This however does not mean that we underestimate other approaches or overestimate the contribution of the systems to be mentioned.

Before we deal with the various speech recognition systems, it will be in order to mention briefly what we mean by speech analysis which forms the preliminaries required to understand the ensuing survey.

The usual way of visual portrayal of speech sounds is the use of sonagraph or spectograms. Spectograph (or Analyzers) produces a permanent spectogram of a complex signal. Developed at the Bell Telephone Labs [15] in mid-forties, it is used in extracting many speech features such as formant structure, voicing, friction, nasality and pitch.

2.1 Hardware Systems

An example of the use of hardware system in the recognition of conversational speech is given by Suzuki[30] usin a 26-channel spectrum analyzer covering frequency range from 200 Hz to 5,900 Hz. The channel outputs were separately grouped and connected to individual vowel decision circuits. The speech was segmented into voiced and unvoiced segments and envelope intensity measurements were also used. The final classification was made by observing the phoneme most frequently recognized.

A device for recognizing Japanese digits spoken in isolation [21] was based on number of voiced interval in each utterance, the presence of an unvoiced consonant and formant frequencies and frequency slopes at three time points in each voiced interval.

Another promising approach to real time speech recognition using analog threshold logic (ATL) gates has been developed and the RCA Laboratories [22]. These gates give zero output until sum of excitatory and inhibitory inputs is less than a specified threshold. Once this condition is satisfied, the output is linearly proportional to the sum. The basic ATL gates are used to abstract relevant features from the output of the spectrum analyzer. Some of the various features extracted are formants, zeros positive and negative slopes etc. Such a feature detector, instead of a categorical decision as to whether a feature is present or absent, gives a quantitative measure of the feature. Outputs of these feature detectors are combined in appropriate manner to perform recognition.

A system for recognizing Italian numerals [11] uses a 17 channel spectrum analyzer followed by threshold detectors which reduce the range of outputs from each channel to binary statements. These signals are then sampled and connected to separate circuits for recognition of each digit.

2.2 Hardware-Computer Combination

Denes [6] gives a system for recognizing the ten digits from complete word patterns. Outputs of a 17 channel spectrum analyzer were sampled and recorded on a magnetic tape. This is then inputted to a computer to form a time-frequency patterns for number of utterances of each word and stored as reference. Cross-correlation process was used to compare the unknown utterance with the stored patterns. Best match pattern was chosen as that of the unknown signal.

Another system due to Gold [13] uses a 16 channel spectrum analyzer, a pitch extractor and a voicing detector. The computer program segmented the words into approximate phonemic units on the basis of voicing magnitude and spectral information. Further spectral durational and intensity measurements were then made on a group of five segments centered on one segment which was defined as stressed. A scoring method, based on the similarity with previous measurements on known words was used to classify the word.

2.3 Adaptive Recognizers

The system is trained first for an utterance by a single talker and then used for recognition purposes. A recognition system [5] using adaptive threshold elements incorporates a 15 channel spectrum analyzer, a 13 level amplitude quantization, and the output is sampled every ten m.sec. This is then fed to ten adaptive threshold elements simulated by computer.

Another system [9] divides the frequency range of 300 Hz to 3000 Hz into ten channels, each of which is envelope detected. Two more signals are formed from the energy above 3 KHz and the overall spectrum. These 12 signals are sampled for $\frac{1}{2}$ sec. after the onset of a word and the resulting 192 samples were punched on to a paper tape to be fed to computer subsequently. Word recognition is by means of linear decision hyperplanes which are adjusted to give optimum performance during a training session.

Recently Clapper [2] has given a scheme for adaptive memory technique to adjust to each speaker. Separately spoken individual words can be automatically recognized using a two dimension pattern of spectral density vs. a nonlinear time base. The pattern for a given word differs from person to person and must be adaptively learned by the machine for each speaker. Simple circuitry is given that learns a word with a single utterance and recognizes it thereafter.

2.4 Usual Accoustic Analysis

A highly successful recognition system based on direct analysis of speech wave has been developed by Reddy [27]. Direct processing of speech wave for segmentation and feature extraction is the significant aspect of this system. Speech wave after normalization is subjected to a segmentation procedure which identifies sustained and transitional segments. The main parameters used for this segmentation are intensity and zero crossing

rate of consequetive 10 msec. sections of the signal. Minimal segments (10 msec.) are grouped together if they are to be accoustically similar, similarity being defined on the basis that intensities of adjacent elements does not vary by more than a specified tolerance interval.

A vowel recognition system [30] performs the spectral analysis by means of an 'analogue ear'. The output of analogue ear was sampled at a high rate in 4 msec. bursts synchronized with the onset of each glottal cycle. Analysis was performed by a computer by means of correlation operations.

Apart from these studies excepting for one study on intonation analysis [10] other study of intonation pattern exists only from the psychological point of view. For example from the perceptual point of view the relation between the movement of fundamental frequency f_0 with respect to time in giving rise to various intonation pattern has been brought out in [12]. This does not involve any recognition by a machine but throws much light on the perception of the intonation patterns.

Thus, we realize the necessity or desirability of using intonational analysis through a computer system which does not exist elsewhere. We describe our approach in this regard in the ensuing section .

CHAPTER 3

STATEMENT OF PROBLEM AND PROPOSED SOLUTION

It has been observed from the literature survey on the speech recognition techniques that very little significant work has been done in the field of intonation analysis. Total continuous speech carries important information in the form of intonation contour. Hence the need arises of the analysis, recognition and classification of intonation patterns which would be a step towards the enhancement of the work on speech recognition.

The aim of this thesis is to obtain a procedure to be implemented on digital computer to analyse and recognize the intonation patterns of the specified sentences.

The variation of the intonation contour is essentially a plot of fundamental frequency f_o of the speech signal with respect to time. f_o varies during the utterance of the sentence, hence its evaluation versus time by digital computation constitutes the crux of the problem. It has been mentioned in Chapter 1 that f_o lies in the range of 70 to 500 Hz. for male, female and children speeches. However, for our purposes, we would limit this range from 70 to 300 Hz which includes the f_o signals of male and female speakers.

In order to process the speech signal in the digital computer, it is essential to have the analog speech signal converted to digital signal. An A/D converter would give the corresponding digital signal for speech. The extraction of the information about f_o at various time intervals can be done by having a bank of band pass filters, covering the entire range of 70 to 300 Hz. It would be desirable to have the bandwidth of filters as narrow as possible, hence giving rise to a large number of filters in the bank. Compromise between the bandwidth and the number of filters would have to be done depending upon the filters used. We would be using the digital filters whose stability is a factor of the word length of the computer, the amount of round off error, and truncation error and the location of the poles of transfer function with respect to the unit circle in z domain [25]. Therefore the number of filters and the bandwidth chosen would be decided upon by:

(a) Stability of the filters.

(b) Time requirement for computation

(c) Memory requirements of the program.

The system which would be used for intonation analysis-recognition is shown in Figure 3.1. The following assumptions are made in the analysis.

(a) The continuous signal to be analysed has the explicit boundaries (of the sentence) defined already.

(b) The duration of the sentence spoken is approximately of 2.4 seconds.

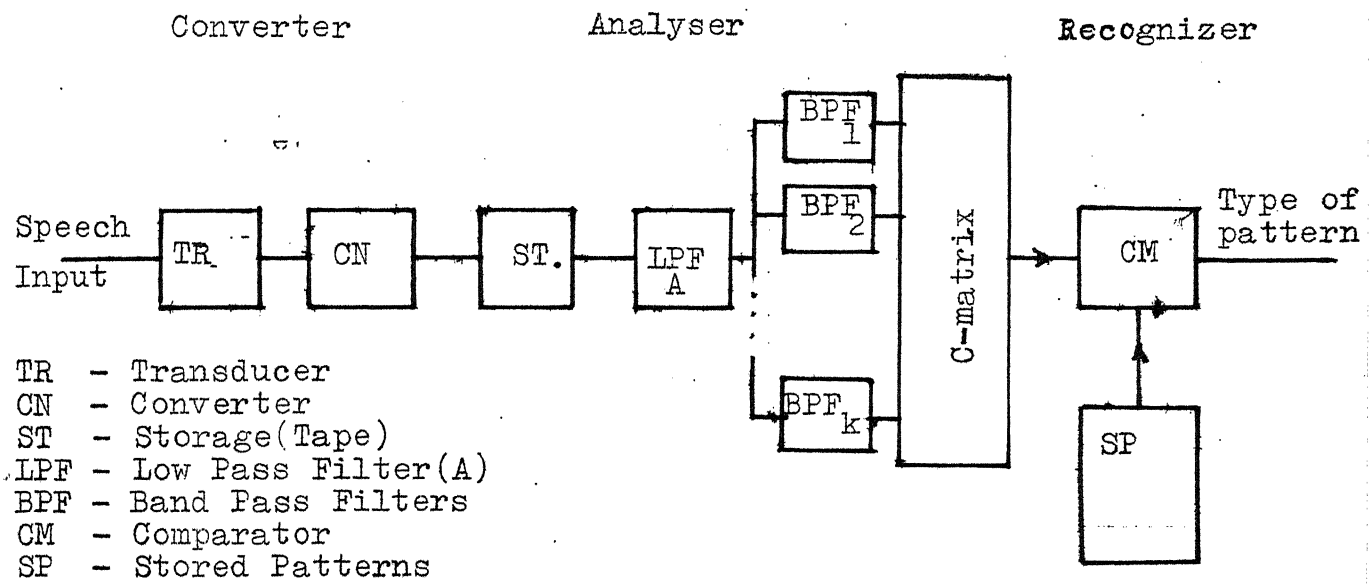


Figure 3.1: Proposed analysis recognition system for intonation patterns.

The system comprises of three main blocks:

(a) Converter

(b) Analyzer

(c) Recognizer.

The converter constitutes a transducer and an A/D converter. The transducer produces appropriate electrical signal corresponding to input speech, to be fed to the A/D converter. The A/D converter gives the digital output for the analog (speech) signal input. The speech signal would be sampled at chosen regular intervals and a binary coded signal would be available corresponding to these sampled values. The sampling should be done

at a rate twice the maximum frequency present in the speech. The output of the converter is stored on a magnetic tape to be processed on the computer later on.

The blocks analyser and recognizer are simulated within the computer. The digitized speech is first passed through the low pass filter 'A' to give signal sequence containing frequencies below 500 Hz. This sequence is then passed through the band pass filters BPF_1 to BPF_k successively. The output of the filter bank is stored in the C matrix (code matrix), which represents the time-frequency relation of the speech signal. This is the intonation pattern of the utterance or the speech signal.

A new set of computational procedure is required by which the contents of C-matrix can be classified by the recognizer. The standard intonation patterns for various types of sentences are stored in the computer memory. By selecting suitable algorithms the C matrix contents should be compared with the stored pattern. Decision on the type of the pattern should be made based on the above comparison. This aspect of the system requires a further investigation.

However, in this thesis we are restricting ourselves only to the analyzer part and rest of the system is assumed. In the next chapter, we give the details of a simulated analyzer.

CHAPTER 4

SIMULATED ANALYZER

Processing of the digitized speech can best be analysed for various frequencies by passing the signal through a bank of digital filters simulated in the computer. What follows in this chapter is a discussion on realization forms of digital filter transfer function, the design criterion, simulation of filter bank and the choice of these filters for our purposes.

The term digital filter refers to a system which executes an algorithm by which a sampled signal or sequence of numbers x(nT), acting as an input is transformed into a second sequence of numbers y(nT) termed the output signal.

One such algorithm used very often is

$$y(nT) = \sum_{i=0}^{N} a_i x(nT-iT) - \sum_{i=1}^{M} b_i y(nT-iT) \qquad (1)$$

Knowing the present and (N-1) past values of input sequence x(nT) and the M past values of output sequence y(nT) the present value of the output can be computed. The difference equation (1) in time domain gives the transfer function of the digital filter in z domain as

$$H(z) = \frac{\sum_{i=o}^{N} a_i \; z^{-1}}{1+ \sum_{i=1}^{M} b_i z^{-1}} \qquad (2)$$

where z^{-1} represents the unit delay operator.

The transfer function H(z) given by (2) leads to two types of filter realizations: (a) where at least one a_i and one b_i is nonzero - gives recursive digital filter. and (b) where all b_i are zero, which gives nonrecursive or transversal digital filter.

Recursive digital filter output depends not only on the input sequence but also on the previous values of the output. Whereas the transversal filter output depends entirely on the input sequence. Both these types have distinct characteristics. The transversal filter possesses excellent phase characteristics, but it requires very large number of terms (hence more number of multiplication and add operation to obtain a sharper cut off. However the recursive filter requires few terms to give a much sharper cut off. Thus to obtain same cut off recursive filter requires less computational efforts.

The required transfer function H(z) for digital filter can be obtained in one of the two ways: (a) obtain the transfer function in s-domain for the corresponding analog filter and apply suitable transformation to obtain

z-domain transfer function, [19], (b) obtain the transfer function of the digital filter directly in z-domain [25]. Both methods have their advantages. In the first case, the use could be made of the extensive work done in analog filter synthesis. Among the various choice of filters, a few are Butterworth, Chebychev and elliptic filters. All these filter characteristics are approximations to a desired rectangular pass band. The Butterworth kind achieves this through monotonic amplitude versus frequency characteristics. By allowing a ripple in pass-band the Chechychev kind using the same number of poles and zeros can achieve sharper cut off. Elliptic filters yield even sharper cutoff than Chebychev for same network complexity.

Synthesis techniques for analog filters yield transfer function in s-domain:

$$H(s) = \frac{\sum_{k=o}^{K} c_k s^k}{\sum_{l=o}^{L} d_l s^l} \qquad K < L \qquad (3)$$

The transfer functionoof digital filter can be obtained from this by applying one of the following transformations:

(a) Standard z-transformation

(b) Bilinear z-transformation.

In first case transfer function H(z) is evaluated using the transformation $s \rightarrow \frac{\log_e z}{T}$, where T is sampling frequency.

Thus

$$H(z) = H(s)\Big|_{s=\frac{\log_e z}{T}} \qquad (4)$$

This is conveniently done by breaking H(s) into partial fraction and then making the substitution

$$\frac{1}{s+a} \rightarrow \frac{T}{1-z^{-1}e^{-aT}}$$

Standard z transformation is good in preserving the impulse response but its frequency response suffers from the aliasing error. This error is due to the fact that H(s) is not band limited. However, if it is band limited, i.e. $H(s) = 0$ for $|w| > w_s/2$; w_s being angular sampling frequency, then aliasing error would be absent.

This aliasing error, can however be eliminated by using the bilinear z transformation or z-form. This requires the substitution

$$s \longrightarrow \frac{1-z^{-1}}{1+z^{-1}}$$

This transformation, eliminating the aliasing error, carries a nonlinear warping of the frequency scale due to the fact that the whole of the frequency scale axis of s plane is compressed into a portion $|w| \leq w_s/2$. This can be compensated by prewarping the critical frequencies of the analog filter in such a way that application of z-forms will shift these frequencies back to their original values.

Thus

$$W_{A_i} = \tan \frac{w_{D_i}}{2} \qquad (5)$$

gives the analog frequency for corresponding digital frequencies w_{D_i} [25].

Another transformation matched z transformation could also be used for obtaining H(z) [14].

It is usual practice to design a low pass filter first and then using suitable transformations, high pass, band pass or band stop filters could be obtained. These transformations could be applied either in analog filter [31] or in digital filter [3] transfer function.

It has been shown [3] that if H(z) is the transfer function of low pass filter, then H(1/f(z)), where

$$f(z) = \frac{z^{-1}(z^{-1}-\alpha)}{(1-\alpha\ z^{-1})}$$

acquires a band pass characteristics.

Here α is so chosen as to give appropriate cut off of the band pass filter. α is given by

$$\alpha = \cos(2k\ w_o) = \frac{\cos k(w_2 + w_1)}{\cos k(w_2 - w_1)} \qquad (6)$$

where w_o = centre frequency of band pass filter

w_2 = upper cut off frequency of BP filter

w_1 = lower cut off frequency of BP filter

$k = \frac{\pi}{\Omega_s}$; Ω_s = angular sampling frequency.

The cut off frequency for the corresponding low pass filter is given by

$$w_c = w_2 - w_1 \tag{7}$$

Realization of Filters

A recursive filter represented by (1) can be realized (or simulated on computer) using either direct, cascade or parallel form as shown in Figure 4.1.

The direct form is the direct realization of the expression of transfer function whereas cascaded form is obtained by expressing transfer function as

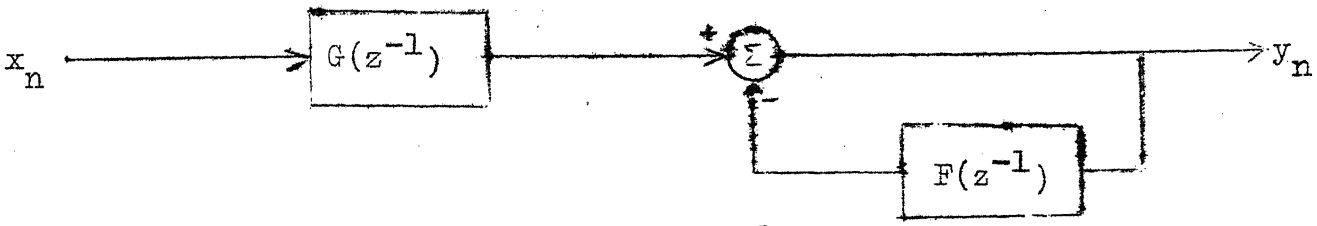
$$H(z) = H_1(z^{-1}) \times H_2(z^{-1}) \times \ldots H_k(z^{-1})$$

where each of the subfilters includes a subset of the poles and zeros of $H(z^{-1})$. Parallel form is obtained by expressing $H(z^{-1})$ in its partial fractions.

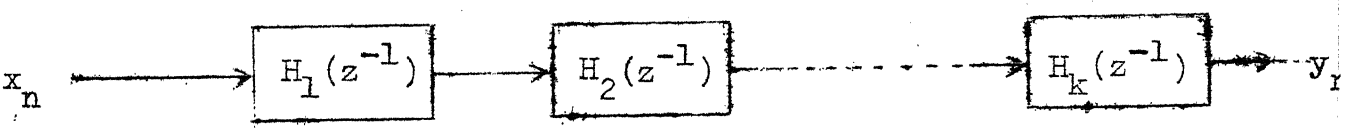
Direct form requires a larger amount of memory and greater accuracy in filter parameter determination. This is seldom preferred to the cascade or parallel realizations. However, cascade form is preferred over parallel form because the errors due to finite word length in computer is pronounced in parallel form.

System Design

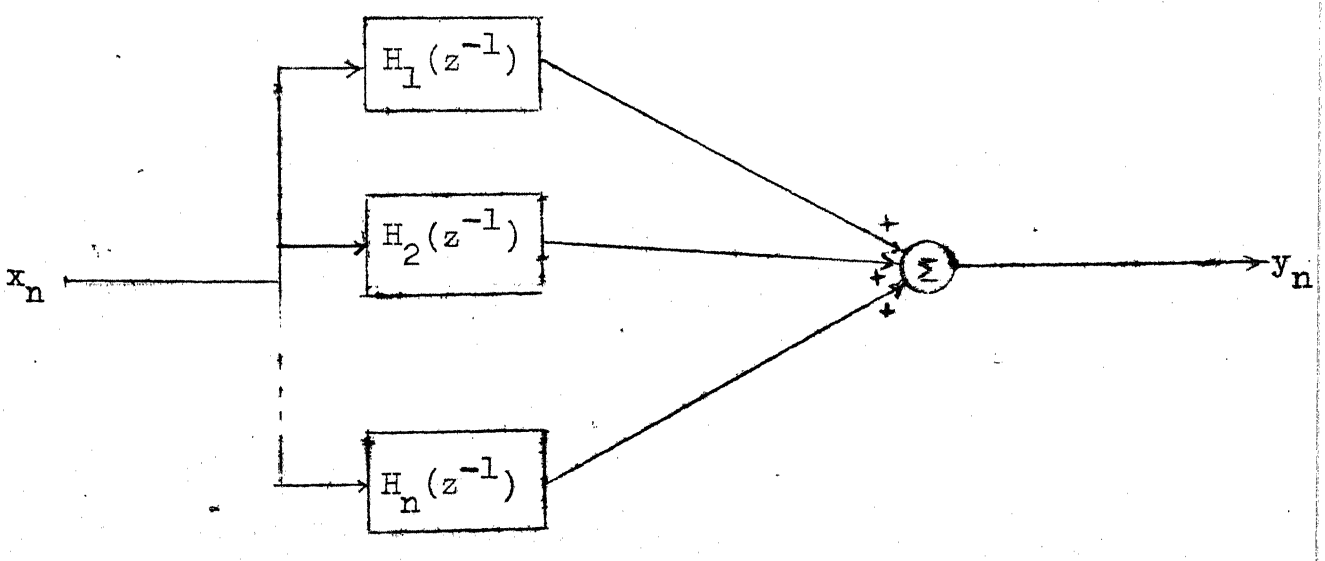
The speech output of the A/D converter has been put on the magnetic tape, for processing later on the computer.



Direct form $H(z^{-1}) = \dfrac{G(z^{-1})}{1+F(z^{-1})}$



Cascade form $H(z^{-1}) = H_1(z^{-1}) \times H_2(z^{-1}) \ldots H_k(z^{-1})$



Parallel form $H(z^{-1}) = H_1(z^{-1}) + H_2(z^{-1}) + \ldots + H_n(z^{-1}$

Figure 4.1: Various realizations of digital filter.

The sampling rate of the signal is 20 KHz using a word length of 12 bits. Offset binary was used in coding.

As has been pointed out earlier, the intonation contour is the variation of fundamental frequency f_o, of speech with time. It has also been shown that this fundamental frequency f_o of speech lies in the range of 70 to 500 Hz. 70 Hz to 160 Hz for male speech, upto 300 Hz for female speech and upto 500 Hz for children. Therefore, in order to evaluate the intonation contour, it becomes necessary that the signal be passed through a low pass filter with cut off at 500 Hz. before processing. The resulting signal could then be analysed for intonation contours. Specifications for this filter would be

Cut off frequency	500 Hz
Pass band tolerance	25 percent
Stop band tolerance	15 percent
Transition ratio	0.8
Sampling frequency	20 KHz.

The type of filter chosen for this is elliptic filter. Since it would give the required filtering with minimum number of cascaded sections. A computer subroutine PROLP due to Sinha [30] was used to carry out this design. This subroutine returns the coefficients of the digital filter which is simulated as subprogram DIGFIL to give the required output.

Filter Bank

The fundamental frequency f_o of 70 to 300 Hz has been divided into contiguous channels of equal bandwidth of 25 Hz each giving the 10 channels of following bands.

70 - 95 Hz	195 - 220 Hz
95 - 120 Hz	220 - 245 Hz
120 - 145 Hz	245 - 270 Hz
145 - 170 Hz	270 - 295 Hz
170 - 195 Hz	295 - 320 Hz

Here again the type of low pass filter used was elliptic filter and the subroutine PROLP was used to obtain low pass filter coefficients. A prototype low pass filter was designed first with following parameters:

Cut off frequency	25 Hz
Pass band tolerance	20 percent
Stop band tolerance	10 percent
Transition ratio	0.9
Sampling frequency	1 KHz

Sampling frequency in this has been chosen as 1 KHz as the maximum frequency present in the signal is 300 Hz. Achieving this, however, is no problem since 1 in 20 samples of the input signal can be taken to suit this condition. Transfer function of the prototype low pass filter is obtained as

$$H(z^{-1}) = \prod_{i=1}^{n} e_i \frac{a_{i1}+a_{i2}z^{-1}+a_{i3}z^{-2}}{b_{i1}+b_{i2}z^{-1}+b_{i3}z^{-2}} \qquad (8)$$

where n is the number of cascaded section.

For further discussions let us take n = 1, which however can be extended to cases n 1 without difficulty.

Converting this low pass filter to band pass filter by the transformation $z^{-1} \rightarrow \frac{-z^{-1}(z^{-1}-\alpha)}{(1-\alpha z^{-1})}$ in (8)

$$H(z^{-1}) = e_1 \frac{a_{11}+a_{12}\left(-\frac{z^{-1}(z^{-1}-\alpha)}{(1-\alpha z^{-1})}\right)+a_{13}\left(-\frac{z^{-1}(z^{-1}-\alpha)}{(1-\alpha z^{-1})}\right)^2}{b_{11}+b_{12}\left(-\frac{z^{-1}(z^{-1}-\alpha)}{(1-\alpha z^{-1})}\right)+b_{13}\left(-\frac{z^{-1}(z^{-1}-\alpha)}{(1-\alpha z^{-1})}\right)^2} \tag{9}$$

Upon simplification, one gets

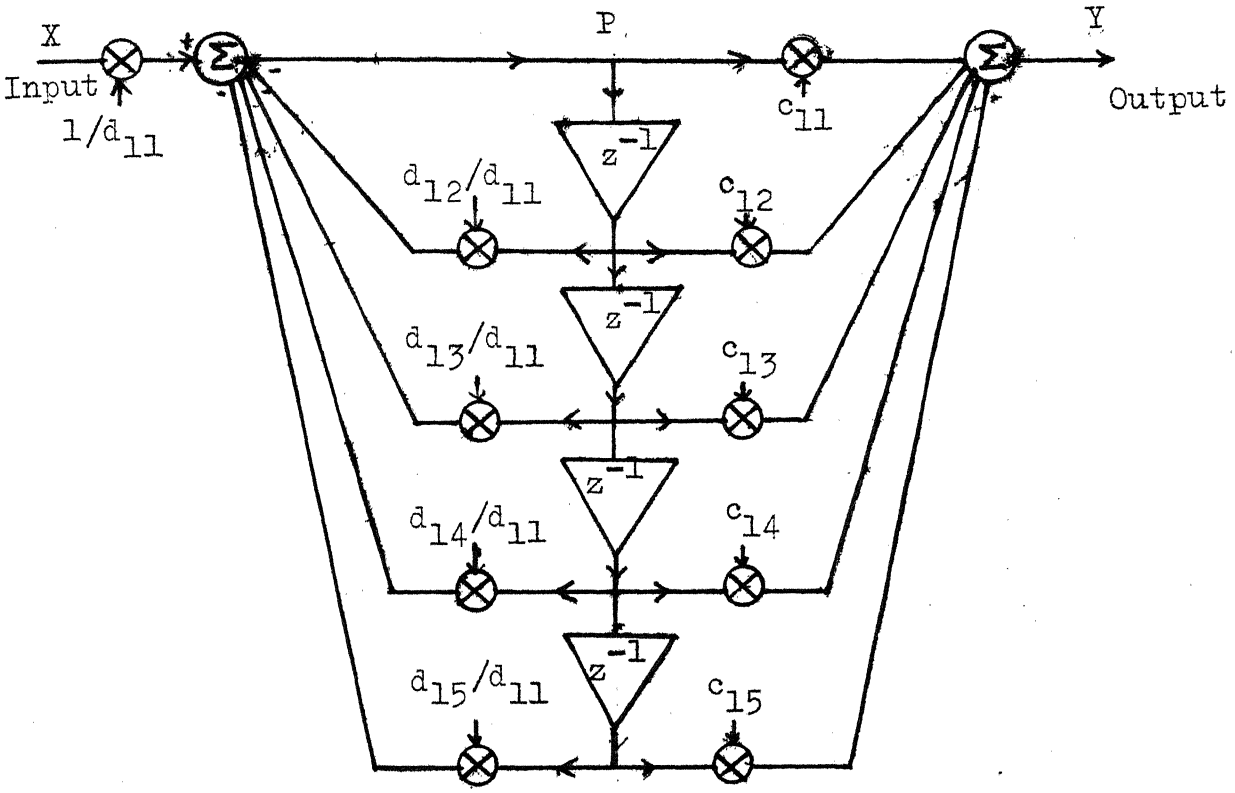
$$H(z^{-1}) = e_1 \frac{c_{11}+c_{12}z^{-1}+c_{13}z^{-2}+c_{14}z^{-3}+c_{15}z^{-4}}{d_{11}+d_{12}z^{-1}+d_{13}z^{-2}+d_{14}z^{-3}+d_{15}z^{-4}} \tag{10}$$

where the coefficients in (10) are related to that of (8) by following relations

$$\begin{aligned}
c_{11} &= a_{11} & d_{11} &= b_{11} \\
c_{12} &= \alpha(a_{12}-2a_{11}) & d_{12} &= \alpha(b_{12}-2b_{11}) \\
c_{13} &= \alpha^2(a_{11}-a_{12}+a_{13})-a_{12} & d_{13} &= \alpha^2(b_{11}-b_{12}+b_{13})-b_{12} \\
c_{14} &= \alpha(a_{12}-2a_{13}) & d_{14} &= \alpha(b_{12}-2b_{13}) \\
c_{15} &= a_{13} & d_{15} &= b_{13}
\end{aligned} \tag{11}$$

Thus one comes to conclusion after observing (11) that the coefficients of various band pass filters can be obtained from a low pass prototype filter coefficients, the only parameter changing is α, which is decided upon by the cut off frequencies of the corresponding filters.

Transfer function in (10) can be represented pictorially as shown below.



z^{-1} represents unit delay.

This has been simulated on digital computer using the subprogram DIJFIL. The unit delay z^{-1} represents storage for one data sample.

<u>Testing of Filters</u>

Before using the filters for actual processing, they are tested for (a) Impulse response (b) Frequency response.

Impulse response indicates the stability of the filter. This could be obtained by getting the output of the filter for an input sequence of 1,0,0 ... 0. For a stable

filter the output decays in transient state and becomes zero in steady state.

Frequency response of the filter is obtained by replacing z^{-1} by e^{-jwT} in the expression for transfer function, i.e. from (10)

$$H(e^{-jwT}) = \frac{c_{11}+c_{12}(e^{-jwT})+c_{13}\ e^{-j2wT}+c_{14}\ e^{-j3wT}+c_{15}\ e^{-j4wT}}{d_{11}+d_{12}\ e^{-jwT}+d_{13}\ e^{-j2wT}+c_{14}\ e^{-j3wT}+d_{15}\ e^{-j4wT}} \tag{12}$$

dividing numerator and denominator by e^{-j2wT} and simplifying gives

$$H(e^{-jwT})$$

$$= \frac{[(c_{11}+c_{15})\cos 2wT+(c_{12}+c_{14})\cos wT+c_{13}] + j[(c_{12}-c_{15})\sin 2wT+(c_{12}-c_{14})\sin wT]}{[(d_{11}+d_{15})\cos 2wT+(d_{12}+d_{14})\cos wT+d_{13}] + j[(d_{12}-d_{15})\sin 2wT+(d_{12}-d_{14})\sin wT]} \tag{13}$$

H(w) for various values of w can be plotted to find whether the response is the desired one.

Program Implementation

The aperiodic nature of the speech wave necessitates the creation of a data window. This subset of data (i.e. a portion of speech signal) is passed through the bank filters respectively number of times till the filter output stabilises. It has been found by experimentation that 5 such repetitions give the desired results. The data window chosen was of

50 samples. Once the corresponding output is obtained the r.m.s. value of these 50 samples is taken and compared with the r.m.s. value of the corresponding input samples.

To ascertain whether the output of a particular filter is present or not, use is made of the fact that filter would attenuate the signals outside its pass band. Whereas the signals in the pass band would give larger outputs. Thus if the ratio of the r.m.s. values of the input and output of the filter is above certain threshold then the presence of the particular frequency is indicated in that particular subset of speech signal.

Thus the decision taken at the end of the processing of 50 samples, based on the value ratio will be indicated as 1 or 0. If the signal is above the threshold, indicating the presence of frequency in that period, would be indicated by 1, and the absence by 0.

The next 50 samples are taken subsequently and processed similarly, till all the samples are processed. The signal is then passed through the subsequent filters.

The decisions 1 or 0 at various time instants are stored in c-matrix.

The computer program organization and the flow charts for the above program are given in Appendix B.

CHAPTER 5

RESULTS OF EXPERIMENTS

The low pass filter A (Figure 3.1) for speech signal and the bank of filters were designed and simulated on the computer. The coefficients (i.e. a_i's and b_i's) of the low pass filter A and that of prototype low pass filter are shown in Table 5.1. The coefficients of the various band pass filters were obtained from the low pass prototype filter using equation (11) of Chapter 4. These coefficients (c_i's and d_i's) are tabulated in Table 5.2.

The testing of the simulated analyzer was carried out in the following manner.

A signal comprising of five sinusoids of different frequencies was fed to the program package. Five filters in whose pass bands these sinusoids are covered indicate their presence by 1 in the relevant row. Whereas the absence is indicated by a 0 in the row for which no sinusoid existed. The results are shown in R.1. Another similar signal, but the frequencies of sinusoids staggered from that of the first signal was also fed. This results for this are shown in R.2. These results establish the efficacy of the filter bank.

Because of the unavailability of the digitized speech, signals were synthesised by mixing several

sinusoids together so as to resemble the spectrum of the speech signal. Sinusoids of different frequencies appeared at various time intervals in such a signal. Two such signals were synthesised and were analyzed by the program successively. Results for these are shown in R.3 and R.4 displaying the c matrix. They indicate the presence of frequencies in the various time intervals and in corresponding bands as per the synthesised signal.

It therefore can be inferred that similar type of results would be obtained when a digitized speech signal is given as input to this program package.

Table 5.1

Type	i	a_{i1}	a_{i2}	a_{i3}	b_{i1}	b_{i2}	b_{i3}
Low Pass Filter 'A'	1	162.6208	-320.5499	162.6208	168.6208	-318.3559	159.2280
	2	164.9172	-315.9571	164.9172	205.2250	-316.9883	123.5782
Prototype Low Pass Filter	1	162.6210	-320.550	162.621	168.2081	-318.3559	159.2280
	2	164.9173	-315.9572	164.9173	205.2251	-316.9883	123.5780

Table 5.2

Band Hz	i	c_{i1}	c_{i2}	c_{i3}	c_{i4}	c_{i5}	d_{i1}	d_{i2}	d_{i3}	d_{i4}	d_{i5}
70-95	1	162.620	-562.689	810.841	-562.689	162.620	168.207	-570.513	808.637	-554.865	159.228
	2	164.917	-562.689	806.238	-562.689	164.917	205.225	-633.830	807.269	-491.549	123.578
95-120	1	162.620	-505.554	716.319	-505.554	162.620	168.207	-512.583	714.126	-498.524	159.228
	2	164.917	-505.554	711.721	-505.554	164.917	205.225	-569.470	712.758	-441.637	123.578
120-145	1	162.090	-435.970	614.870	-435.970	162.620	169.207	-442.032	612.676	-429.908	159.228
	2	164.917	-435.970	610.278	-435.970	164.917	205.225	-491.089	611.309	-380.850	123.570
145-170	1	162.620	-355.651	516.414	-355.651	162.620	168.207	-360.596	514.220	-350.705	159.228
	2	164.917	-355.651	511.821	-355.651	164.917	205.225	-400.615	512.853	-310.686	123.578
170-195	1	162.620	-266.574	430.588	-266.574	162.620	168.207	-270.218	428.394	-262.868	159-228
	2	164.917	-266.574	425.996	-266.574	164.917	205.225	-300.277	427.027	-232.872	123.578

Table 5.2 (continued)

Band Hz	i	c_{i1}	c_{i2}	c_{i3}	c_{i4}	c_{i5}	d_{i1}	d_{i2}	d_{i3}	d_{i4}	d_{i5}
195-220	1	162.620	-170.934	365.794	-170.934	162.620	168.207	-173.311	363.600	-168.557	159.228
	2	164.917	-170.934	261.201	-170.934	164.917	205.224	-192.545	362.233	-149.323	123.578
220-245	1	162.620	- 71.085	328.374	- 71.085	162.620	168.207	- 72.073	326.180	- 70.097	159.228
	2	164.917	- 71.085	323.781	- 71.085	164.917	205.225	- 80.070	324.813	- 62.098	123.578
245-270	1	162.620	30.514	321.991	30.514	162.620	168.207	30.938	319.797	30.089	159.228
	2	164.917	30.514	317.398	30.514	164.917	205.225	34.372	318.430	26.656	123.578
270-295	1	162.620	131.362	347.270	131.362	162.620	168.207	133.188	345.076	129.535	159.228
	2	164.917	131.362	342.677	131.362	164.917	205.225	147.970	343.709	114.754	123.578
295-320	1	162.620	228.976	401.737	228.976	162.620	168.207	232.159	399.543	225.792	159.228
	2	164.917	228.976	397.144	228.976	164.917	205.225	257.925	398.175	200.026	123.578

CHAPTER 6

CONCLUSION

The intonation contour is an important feature of speech. Its evaluation would prove to be a significant tool in the continuous speech recognition systems. In the present work a procedure has been developed to find out the intonation contour of an utterance and the feasibility has been established. Some of the specific contributions of the thesis are:

(a) Simulation of bank of digital filters.

(b) Use of frequency transformation techniques to obtain appropriate band pass filters from low pass digital filter transfer functions, for the purpose of speech spectrum analysis.

(c) A procedure to evaluate the intonation contour of a continuous speech signal.

The program package developed, gave satisfactory results with the synthesised speech signals. However, it could not be tested for actual speech signal. The magnetic tape containing the output of A/D converter (by courtsey of T.I.F.R.Bombay) could not be used probably because of the incompatibility of the systems.

Attempts were made to design band pass filters with a narrow band width (10 Hz). This, however, resulted in oscillations at the filter output. It has been found that at narrow bands, the poles of the transfer function become very close to the unit circle. Due to rounding off error in computation, it is possible that the poles fall within the unit circle, leading to instability. This difficulty was, however, overcome by choosing a BW of 25 Hz.

Scope for further work

In order to obtain better intonation patterns, narrow band filters should be used. It is therefore desirable to obtain stable narrow band filters. This particular aspect requires further investigation.

Another aspect which requires further investigation is the selection of suitable algorithms for matching the stored patterns with that of the analysed ones. Once this has been done, the classification of the utterance on the basis of the intonation patterns could be achieved.

APPENDIX A

Speech Recognition and Analysis Review

See References (For work done before 1950):

1961 Marril T., Automatic Recognition of Speech, IRE Trans. on Human Factors in Electronics, HFE 2, No.1, pp. 34-38.

1965 Lindgren, N., Machine Recognition of Human Language, IEEE Spectrum 2,3 and 4.

1965 Flanagan, J.L., Speech Analysis Synthesis and Perception, Springer Verlag, Chap 5.5.

1965 Sposkhkov, M.A., The Speech Signal in Cybernetics and Communication, Joint Pub.Res.Service, JPRS 28, 117, Sec.11.2, 11.4 and 12.2. (Translation).

I. I. T. KANPUR
CENTRAL LIBRARY
Ac. No. A 26282

Year	Scientists	Instruments	Method	Restrictions	Recogn. Score in percent
1952	Davis, Balashek Biddulph (Bell Labs.)	Spoken digit Recognizer-Hardware	Assign the spoken words a most probable digit category use F1/F2 measurements during the vowel segments	Single speaker, 10 digits used, clearly speak, pause between digits.	100 percent digit recognition
1958	''	Improved ''	''	Multiple speakers but only 10 digits and not more vocabulary.	99 percent
1956	Wiren and Stubbs (Northe-Eastern Univ.)	Speech Recognizer using distinctive binary opposition	Distinctive feature binary approach (voiced/voiceless, stop/fricative etc.)	21 talkers	94 percent vowels in short words
1956	Olson and Belar (RCA Labs)	Syallable recognizer + connected typewriter	8 freq.bands, 5 time intervals per syllable (8x5 cells matrix) Store 1 in cell if the signal energy corresponding to that cell is greater, store o otherwise and decode.	10 monosyllables, single speaker, careful pronunciation	98 percent syllable recognition
1961	-do-	Improved -do-	8 freq.bands, 5 time intervals Insert time sample only if significant change in spectral power distribution noticed.	100 syllables	Performance not given.

1959	Denes and Fry (Univ.College, London)	Speech analyzer Spectral pattern matcher and stored phonemic pair probabilities	Used the stored information when spectral pattern matched	4 vowels and 9 consonants	Poor overall performance, but linguistic information increased the recognition by 20 percent.
Hardware and Computer					
1959	Forgie and Forgie (Lincoln Lab)	35 channel filter bank analyzer and a computer	Filter and output is envelope detected, sampled and fed to a computer use F0, F1 and F2 informations already measured from filterbanks	10 vowels in context, 21 talkers (males and females) no adjustments for a talker	93 percent vowel recognized
1961	Suzuki and Nakata (Radio Res.Lab, Tokyo)	Vocoder Type 26 channel spectrum Analyser, output rectifier and smoother circuits; seperate wideband format related channels used	Group channel outputs connect to vowel decision circuits. Estimate F0, voiced/unvoiced, envelope intensity. separate recognition decisions at regular intervals of the voiced sound. Identify most frequent phoneme, or 2 most phonemes	-	-
1962	Sukai and Deshita (Kyote Univ)	Recognizer and separate circuits for segmentation	Use during segmentation the following: measure zero crossings, energy variations at freq. regions. obtain measure of 'stability' of distance between digital patterns generated.	Some phoneme input not allowed. (Details not known)	90 percent vowels, 70 percent consonants

1963	Nagata et al (Nippon Elec.Co.	Japanese Digit Recognizer	Make recognition on the number of voiced intervals in each utterance and presence of unvoiced consonants; Format frequencies format freq.slopes at 3 time points in each voiced interval. 8 such feature detection done. Each features state is indicated by energizing at one of the 56 lines of output, each output connected to resistor matrix rows of r.m.indicate ten digits.Highest voltage in a resister on likely digit.	F0 used for females talker. decision matrix. Single male speaker 1000 utterances of digits (same speaker) 600 utterances, 20 males adjust matrix 10 females, 500 utterances,and adjust matrix.	99.7 percent (digit identifier) 97.9 percent digit identifier 93
1964	Martin et al (RCA lab)	19 channel spectrum analyzer with Q-function:1to 2 AND'gates,mono-stable multi-vibrator and 500 ATL(Analog ThresholdLogic elements)i.e. transistor circuit with excitory/inhinit inputs	Use Neuron network model output only if linear sum of input currents exceed threshold (preset) value (subtract inhibitory inputs).This should be proportional to some of inputs until saturation. Use also various speech features steady state, transition intensity ratios.	No formant peaks or transitions are measured use of spectral regions of increasing/ decreasing energy.(i.e. format-skirt slope)	Primising still going on.

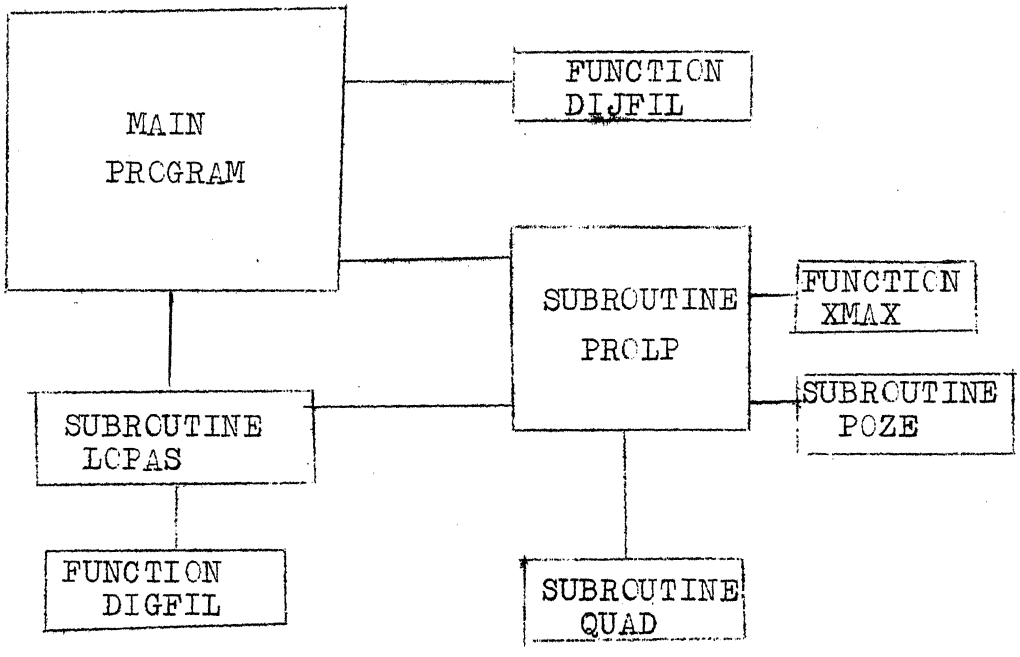
1965	Falter	=	Feature abstractions and weightage assignments for slight variations in the physical characterision of patterns that differ.	CVC utterances, 6 males, 22 different-phoneme-vowels slopes and vowels like sounds	82 percent to 99 percent
1966	Gazdag (Univ. of Ill.)	12 channel spectrum Analyzer (Max. 3KH$_3$ range)	6 Hyper-planes portion the measurement space defined by the outputs of chan.analyzer Implemented by summing amplifiers and trigger circuits. Trigger circuit change when the output of the amplifier passes thru' zero. Final out in 6 binary quantities which signify the time variation and characters of the word spoken	Not completely implemented and system not fully tested. Mainly concern with mathematical investigation of fundamental ideas output in binary no.form.	Promising but time invariant nature of system not provided. No. of result give.
1967/ 1968	Gilli and Meo	17 channel spectrum analyzer. Threshold Detectors CTD	TD reduces the output of channels to binary statements. Sample these signals and connect to recognition circuits. Make decisions on the basis of the sequential occurrence of V+C patterns and transition patterns	10 enumerals 10 speakers Now utterances same speaker crude patterns are enough for digit recognition	100 percent(digit) 90 -do- (Redisigner planned)

1967	Ross (RCA Lab)	Olsen and Belar Word recognizer extension. 4 channel analyzer digit disk storage	Generate 20 bit binary pattern for each word by sampling five 2-state signals derived from the 4 channels analyzer. Mean energy level ratio in each freq.channel to mean energy in overall signal. Compare energy in highest freq. channel w.r.t. 2 lower most channel and thus form 5th signals. Use threshold and compare each signal and produce 5 binary outputs.For the resulting pattern use nearest neighbour concept and compare with stored data.	Experimentor controls the permissible distance bet. same output pattern.Single word presented many time.Add new patterns of stored one. Thus diff. patterns for each word are stored.	5 percent
1960	Dones and Mathews (Bell Lab)	17 channels spectrum analyzer and computer	Output samples recorded on a tape and inputted to a computer. Time freq. patternsfor number of utterances were averaged. and stored for reference. Use cross correlation process for new word comparison. Select best match	1female and 6 males pause between digits. Treat vowel/ consonant time normalisation separately.	94 percent (digit) identifier with time normalization).

1960	Sebestyen (Litton Industry)	18 channel recorder analyzer computer	Computer linear transformation of analyzer outputs and obtain maximal clustering of the data in 361 dimensional space formed by time-sampling the channel outputs. Time normalization was included as one of the space.	400 utterances (10 numerally) 10 talkers (No other detail)	100 percent (digit recognizer) (7 example of each normal)
1962	Meeker et al (RCA Lab)	18 channel recorder analyzer and computer	From channel energy level detect boundary between phoneme. Consider long duration of voice as vowel, if so determine F1 F2 at the centre of the segment.	7200 vowels, from words ten male speaker 10 consonants 6000 words	Phoneme recognition 40 percent 97 percent of consider each talker separately and use vowel character 60 percent.

1970	Dixon and Tappert	12 channel analyzer computer and CRT display	Filter out put rectified at 10 msec intervals. Display digital sound spectrogram/ power spectrum at a given time sample. Average spectrum for a given class, correlation of a given sound class of a given current spectrum with average spectrum. Previous 20 msec period is converted into log form, ditized to 8 bit numbers and sent to the computer. Slowly changing spectral places are marked as boundaries (including transition and boundary values).	Lexical entries used and higher level linguistic constraints. Operator uses function keys and shift the utterences of interest and decide various operations	not evaluated
1972	Das and Stanat (IBM)	Shaping filter system (24 dB/Oct high level) 250/3.3 KHz range only 14 filters used	At 20 msec intervals average filter outputs after each interval 14 filter output energy etc is stored. Spectrographic plot with energy values coded in diff.intensity level used visually and handlabelled. Use several weighted vectors derived in an iterative manner and temporal information stored and used for segmentation (Training of speakers involved).	10 predetermined phrases 20 male speakers 8 utterances 16 males and 14 females	100 percent training set utterences 92 percent (untrained)

APPENDIX B



COMPUTER PROGRAM ORGANIZATION

FUNCTION DIGFIL - computes the low pass filtered sequence.

FUNCTION DIJFIL - computes the band pass filtered sequence

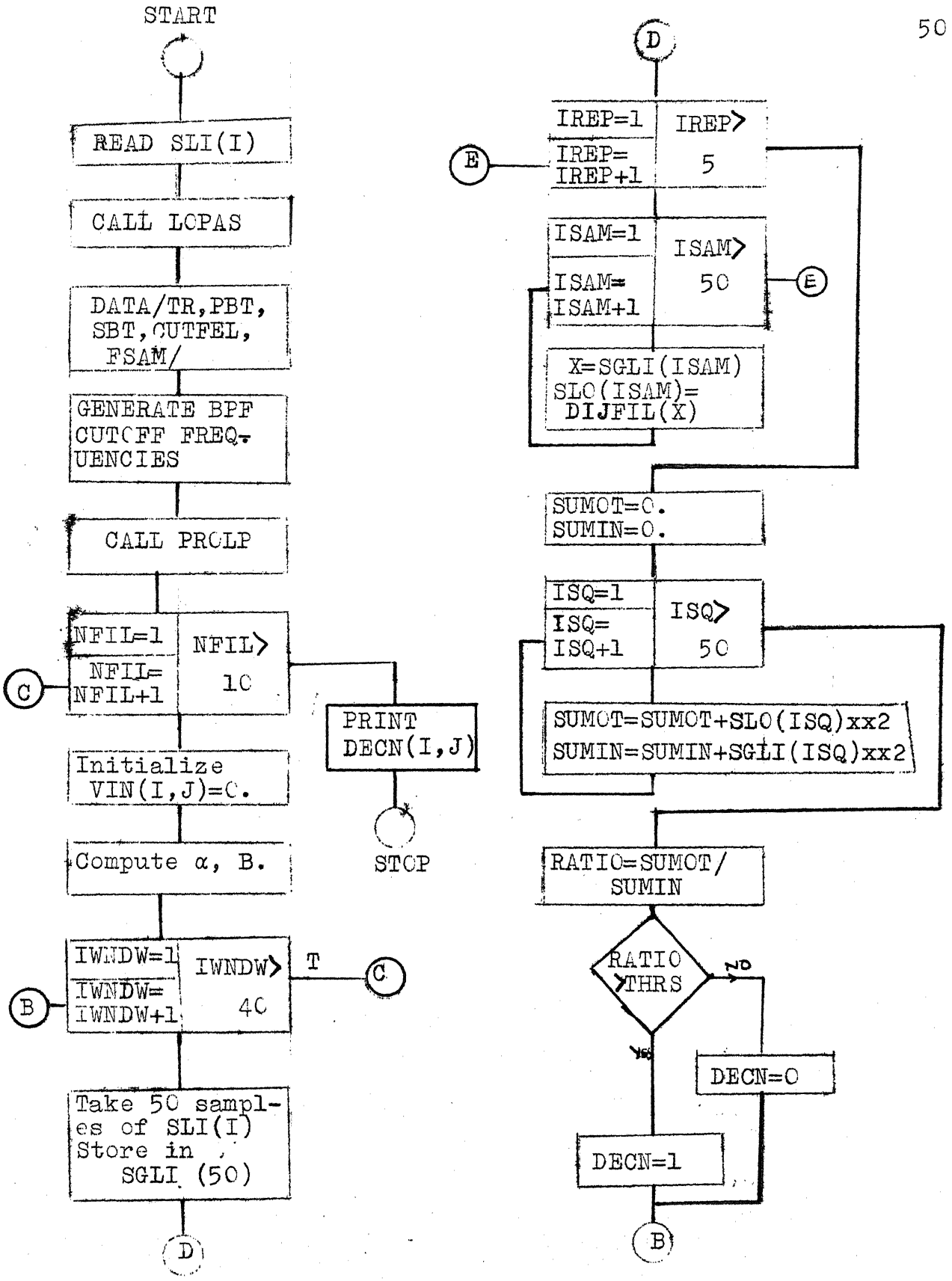
SUBROUTINE LOPAS- performs low pass filtering (for 500 Hz)on the speech input

SUBROUTINE PROLP - computes the coefficients of low pass filter

SUBROUTINE POZE - computes the poles and zeros of the transfer function

SUBROUTINE QUAD - computes the quadratic expressions for transfer function

FUNCTION XMAX - computes the maximum value of an array.



FLOW CHART OF COMPUTER PROGRAM

```
*
C   THIS PACKAGE FINDS THE INTONATION PATTERN OF AN UTTERANCE OF DURAT
C   2.4 SEC.
C   VECTOR SLI(I) CONTAINS SPEECH INPUT SAMPLES.
      DIMENSION A(6,2,3),E(6),SLI(2 00),SGLI(50),SLO(50),VALR(4 )
     ,VIN(6,4),CNT(6),FR( 5),B(6,2,5)
      INTEGER DECN(10,40)
      COMMON VIN
      NTAPE=4
      PI=3.14159
C   READ SPEECH SAMPLES INTO SLI(I) FROM TAPE.
      READ(NTAPE)SLI
C   CALL LOPAS TO ELIMINATE FREQUENCIES ABOVE 500. HZ.
      CALL LOPAS(SLI)
C   DESIGN PROTOTYPE LOWPASS FILTER
C         TR=TRANSITION RATIO.
C         PBT=PASS BAND TOLERANCE.
C         SBT=STOP BAND TOLERANCE.
C         FSAM=SAMPLING FREQUENCY.
C         CUTFFL=CUT OFF FREQUENCY OF PROTOTYPE LOWPAS FILTER( THIS EQ
C                  BAND WIDTH OF THE BP FILTER)
      DATA TR,PBT,SBT,FSAM,CUTFFL/.9,.2,.1,2000.,25./
      PISM=PI/FSAM
C   GENERATE FREQUENCIES FOR BP FILTER
      FR(1)=70.
      DO68I=2,10
      FR(I)=FR(I-1)+CUTFFL
68    CONTINUE
C   SUBROUTINE PROLP RETURNS COEFFCIENTS OF LP FILTER IN ARRAY A(I,J,K
C         I-INDEX OF SECTION CASCADE
C         J-IS 1 FOR NUMERATOR
C         J-IS 2 FOR DENOMINATOR
C         K-1,2,3 DESIGNATE COEFFIECIENTS OF DELAYS
C         NSEC-NUMBER OF SECTIONS IN CASCADE
      CALL PROLP(CUTFFL,TR,PBT,SBT,FSAM,E,A,NSEC)
      K=1
C   COMPUTATION BY FILTER-BANK
C         NO. OF FILTERS=10
      DO70NFIL=1,10
C         FL-LOWER CUTOFF OF BP FILTER
C         FH-UPPER CUTOFF OF BP FILTER
      FH=FR(NFIL+1)
      FL=FR(NFIL)
      ALP=(COS(PISM*(FH+FL)))/COS(PISM*(FH-FL))
C   COMPUTE COEFFICIENTS B(I,J,K) OF BP FILTER,AFTER APPLYING FREQUENC
C   TRANSFORMATIONS TO LP FILTER.
      DO25I=1,NSEC
      DO25J=1,2
      B(I,J,1)=A(I,J,1)
      B(I,J,2)=ALP*(A(I,J,2)-2.*A(I,J,1))
```

```
      B(I,J,3)=(A(I,J,1)-A(I,J,2)+A(I,J,3))*ALP**2-A(I,J,2)
      B(I,J,4)=ALP*(A(I,J,2)-2.*A(I,J,3))
      B(I,J,5)=A(I,J,3)
35    CONTINUE
      DO35I=1,NSEC
      DO35J=1,2
      DO35 K=2,5
      B(I,J,K)=B(I,J,K)/B(I,J,1)
      CNT(I)=B(I,1,1)/B(I,2,1)
35    CONTINUE
      DO36I=1,NSEC
      DO36J=1,2
      B(I,J,1)=1.
36    CONTINUE
C  INITIALISE VIN(I,J)=0.
      DO71I=1,6
      DO71J=1,6
      VIN(I,J)= .0
71    CONTINUE
      NB=1
      NO=1
      DO72 IWNDW=1,38
C  READ 50 SAMPLES OF INPUT INTO SGLI(I).
      DO73J=1,50
      SGLI(J)=SLI(NB)
      NB=NB+1
73    CONTINUE
      DO74 IREP=1,5
C  OUTPUT OF BP FILTER IS IN ARRAY SLO(I).
      DO74ISAM=1,50
      X=SGLI(ISAM)
      SLO(ISAM)=DIJFIL(X,NSEC,B,E)
74    CONTINUE
C  COMPUTE RMS VALUE OF INPUT AND OUTPUT SEQUENCES.
      SUMIN=0.0
      SUMOT=0.0
      DO95ISQ=1,50
      SUMIN=SUMIN+SGLI(ISQ)**2
      SUMOT=SUMOT+SLO(ISQ)**2
95    CONTINUE
C  EVALUTE RATIO OF THE RMS VALUES OFTHE TWO SEQUENCES.
      RATIO=SQRT(SUMOT/SUMIN)
      VALR(IWNDW  )=RATIO
C  COMPARE RATIO WITH THRESHOLD VALUE THRS.
      IF(RATIO.GT.THRS) GOTO94
      DECN(NFIL,IWNDW+1)=0
      GOTO72
94    DECN(NFIL,IWNDW+1)=1
72    CONTINUE
      PRINT 104,(VALR(IK),IK=1,38)
```

```
10    CONTINUE
C  OUTPUT THE C-MATRIX,
C         J-TIME COUNTER,ONE INCREMENT REPRESENTS 50 MSEC. FOR 2KHZ.
C         I- BP FILTER COUNTER.
      PRINT 217
      PRINT219
      PRINT218
      DO92IM=1,20
      K=21-IM
      IL=K
      IH=K+1
      PRINT103,FR(IL),FR(IH),(DECN(K,J),J=1,38)
92    CONTINUE
      PRINT220
      PRINT221
102   FORMAT(X,10E10.3)
104   FORMAT(X,19F6.1)
218   FORMAT(X,*FILTER BAND*/)
220   FORMAT(/,63X,*TIME -----*,)
103   FORMAT(1H0,2F5.0,2X,39I3)
219   FORMAT(1 (/))
221   FORMAT(///,25X,* R.A C-M A T R I X  F O R  A N I N T O N A T I
     2 P A T T E R N*)
115   FORMAT(X,19F6.2,4X,2F6.1)
217   FORMAT(1H1)
      STOP
      END
```

```
*IBFTC
      FUNCTION DIJFIL(X,N,B,E)
C  COMPUTES THE BP FILTERED OUTPUT OF THE INPUT SEQUENCE.
C  VIN(I,J)- TEMPORARY STORAGE FOR SEQUENCE
C         J- REPRESENTS DELAY NUMBER
      DIMENSION B(6,2,5),E(6),VIN(6,4)
      COMMON VIN
      DO1I=1,N
      P=(X-B(I,2,2)*VIN(I,1)-B(I,2,3)*VIN(I,2)-B(I,2,4)*VIN(I,3)-
     1B(I,2,5)*VIN(I,4))/B(I,2,1)
      X=(B(I,1,1)*P+B(I,1,2)*VIN(I,1)+B(I,1,3)*VIN(I,2)+B(I,1,4)*VIN
     2)+B(I,1,5)*VIN(I,4))*E(I)*2.293692/1.527197
      VIN(I,4)=VIN(I,3)
      VIN(I,3)=VIN(I,2)
      VIN(I,2)=VIN(I,1)
      VIN(I,1)=P
1     CONTINUE
      DIJFIL=X
      RETURN
      END
```

```
*
      SUBROUTINE PROLP (OM1,TR,R1,R2,OMS,E,A,NT)
C  COMPUTES THE COEFFICIENTS OF THE CASCADED TRANSFER FUNCTION.
      DIMENSION E( 6),A(6 ,2,3),GAMA(21),ZERO(21),UP(21),XP(21)
      COMPLEX POLE(2 ),GUM
      EP=SQRT((R1*2.)-(R1**2))/(1.-R1)
      AL=SQRT(1.-R2*R2)/(EP*R2)
      OM2=OM1/TR
      XK=1./(2.*OMS)
      CON=TAN(6.28318*XK*OM1)
      DON=1./CON
      GON=DON*DON
      GA=AL
      I=0
      I=I+1
      GAMA(I)=GA
      GA=(GA+1.)/(2.*SQRT(GA))
       ROM2=(OMS/3.14159)*ATAN(CON*GA)
      IF(ROM2.GT.OM2)GOTO 1
      TR=OM1/ROM2
      NT=(2**I)/2
       CALL POZE(EP,GAMA,I,ZERO,UP,XP)
      CALL QUAD(DON,GON,NT,ZERO,UP,XP,E,A)
      J=0
      XE=XMAX(E,NT)
      J=J+1
      IF(E(J).EQ.XE)GOTO 3
      GOTO 2
      E(J)=E(J)*(1.-R1)
      RETURN
      END
```

```
*IBFTC
      SUBROUTINE POZE(EP,GAMA,NT,ZERO,UP,XP)
C  COMPUTES THE POLES AND ZEROS OF THE TRANSFER FUNCTION.
      COMPLEX POLE(20),GUM
      DIMENSION GAMA(21),ZERO(20),UP(20),XP(20)
      X=1./EP
      ZERO(1)=SQRT((1.+GAMA(1))/2.)
      POLE(1)=CMPLX(0.,X)
      X=ZERO(1)
      GUM=CMPLX(X, .)
      X=GAMA(1)
      POLE(1)=GUM*CSQRT((CMPLX(1.,0.)+POLE(1))/(CMPLX(X,0.)+POLE(1)))
      IF(NT.EQ.1)GOTO 5
      DO 1 I=2,NT
      NJIG=(2**I)/2
      X=GAMA(I)
      GUM=CMPLX(X, .)
      DO1 J=1,NJIG
      NG=NJIG+J
      ZERO(NG)=SQRT(((1.+X)/2.)*(1.-ZERO(J))/(X-ZERO(J)))
      ZERO(J)=SQRT(((1.+X)/2.)*(1.+ZERO(J))/(X+ZERO(J)))
      POLE(NG)=CSQRT((CMPLX(1.,0.)+GUM)*(CMPLX(1.,0.)-POLE(J))
     /(CMPLX(1.,0.)*(GUM-POLE(J))))
      POLE(J)=CSQRT((CMPLX(1.,0.)+GUM)*(CMPLX(1.,0.)+POLE(J))
     1/(CMPLX(1.,0.)*(GUM+POLE(J))))
      CONTINUE
      NJIG=(2**NT)/2
      DO2 J=1,NJIG
      UP(J)=REAL(POLE(J))
      XP(J)=AIMAG(POLE(J))
      CONTINUE
      RETURN
      END
```

```
*IBFTC
      SUBROUTINE QUAD(DON,GON,NT,ZERO,UP,XP,E,A)
C  COMPUTES THE QUADRATIC EXPRESSION FOR FILTER TRANSFER FUNCTION
      DIMENSION ZERO(20),UP(20),XP(20),E(6),A(6,2,3)
      DO1 I=1,NT
      A(I,1,1)=GON+ZERO(I)**2
      A(I,1,2)=-2.*(GON-ZERO(I)**2)
      A(I,1,3)=A(I,1,1)
      YY=UP(I)**2+XP(I)**2
      ZZ=2.*ABS(XP(I))*DON
      A(I,2,3)=YY-ZZ+GON
      A(I,2,2)=2.*(YY-GON)
      A(I,2,1)=YY+ZZ+GON
      E(I)=(A(I,2,2)+A(I,2,2)+A(I,2,3))
     1/(A(I,1,1)+A(I,1,2)+A(I,1,3))
      CONTINUE
      RETURN
      END
```

```
      FUNCTION XMAX(COL,N)
      DIMENSION COL(6)
      XMAX=0.
      DO11 I=1,N
      IF((ABS(XMAX)-ABS(COL(I))).GE.0.)GO TO 11
      XMAX=ABS(COL(I))
   11 CONTINUE
      RETURN
      END
```

```
*IBFTC
      SUBROUTINE LOPAS(SLI)
C  SUBROUTINE COMPUTES THE LOWPASS FILTERED OUTPUT.
      DIMENSION SLI(2000),C(6,2,3), (6),RIN(6,2)
      COMMON RIN
      DATA TRL,PBTL,SBTL,FSAML,CUTFFL/.8,.25,.15,20000.,500./
      DATA RIN/12*0./
      CALL PROLP(CUTFFL,TRL,PBTL,SBTL,FSAML,B,C,N)
      DO1I=1,2000
      X=SLI(I)
      SLI(I)=DIGFIL(X,N,B,C)
1     CONTINUE
      RETURN
      END
```

```
*IBFTC
      FUNCTION DIGFIL(V,N,L,A)
      DIMENSION A(6,2,3),E(6),RIN(6,2)
      COMMON RIN
      DO 2 I=L,N
      P=(V/A(I,2,1))-(A(I,1,2)/A(I,1,1))*RIN(I,1)
     1-(A(I,2,3)/A(I,2,1))*RIN(I,2)
        V=E(I)*(A(I,1,1)*P+A(I,1,2)*RIN(I,1)
     1+A(I,1,3)*RIN(I,2))*1.19692/1.527197.0
      RIN(I,2)=RIN(I,1)
      RIN(I,1)=P
    2 CONTINUE
      DIGFIL=V
      RETURN
      END
```

```
C   THIS PROGRAM READS THE TAPE
C   TAPE TO BE MOUNTED ON B-4
C   ARRAY ADOUT CONTAINS ONE RECORD LENGTH
C   ARRAY DAT STORES EACH OF 12 BITS.
C   ARRAY PAT CONTAINS DECIMAL EQUIVALENT OF THE12 BITS CORRESPONDIN
C     ONE SAMPLE
C         IRCORD=NO. OF RECORDS TO BE READ
       DIMENSION PAT(4000)
      INTEGER ADOUT(1333),DAT(4000),A,B,C,D,E
      IRCORD=10
      DO20IM=1,IRCORD
      DO1I=1,1333
1     ADOUT(I)=0.0
      E=2**12
C   READ THE TAPE
      CALL READ(ADOUT)
      IFLAG=0
      J=1
      DO10 I=1,1333
C   SEPERATE THE 36 BITS INTO 12 BITS EACH
      A=ADOUT(I)
      IF(A.GE.0) GOTO5
      IFLAG=1
      A=-A
5     C=A/E
      B=A-C*E
      D=C/E
      C=C-D*E
      IF(IFLAG .NE. 0) D=D+E/2
      DAT(J)=D
      DAT(J+1)=C
      DAT(J+2)=B
      J=J+3
      IFLAG=0
10    CONTINUE
      REWIND 1
      READ(1) DAT
C   CONVERSION TO DECIMAL
      DO 15K=1,4000
      PAT(K)=DAT(K)
      PAT(K)=10.*(PAT(K)/4096.)
15    CONTINUE
      PRINT 103,PAT
20    CONTINUE
100   FORMAT(50X,I7/)
101   FORMAT(5X,10O12)
103   FORMAT(5X,20F6.2)
      STOP
      END
```

```
*IBMAP CONV
C       MAP SUBROUTINE TO READ THE TAPE
        ENTRY   READ
READ    SAVE    1,2,4
        CLA     3,4
        STA     IO
        ENB     =0
        TCOB    *
        RDS     TB4
        RCHB    IO
        TRCB    ERR
        ENB     =0000000000003
        RETURN  READ
IO      IORD    **,,1333
TB4     BOOL    2224
ERR     TRA     S.JXIT
        END
```

BIBLIOGRAPHY

1. Atal B.S., 'Automatic speaker recognition based on pitch contour', JASA, Vol.52, No.6, Dec.1972, pp.1687-1697.

2. Clapper G.L., 'Automatic word recognition', IEEE Spectrum, Aug. 1971, pp. 57-69.

3. Constantinides A.G., 'Frequency transformations for digital filters', Electronic Letters, Vol. 3, March 1967.

4. Cannon M.W., 'A method of analysis and recognition for voiced vowels', IEEE Trans. on Audio and Electroacoustics, 1968, AU-16, No.2, pp. 154-159.

5. Dammann J.A., 'Application of adaptive threshold elements to recognition of acoustic-phonetic states', JASA, Vol.38, 1965, pp. 213-223.

6. Denes, P. et.al., 'Spoken digit recognition using time frequency pattern matching', JASA, Vol.32, Nov. 1960, pp. 1450-1455.

7. Fant G., Accoustic Theory of Speech Production, Mouton and Co., 1960.

8. Flanagan, J.L., Speech Analysis, Synthesis and Perception, Springer-Verleg, 1972.

9. Fraipont, D. 'Voice actuated address mechanism', Report No.3, Electronic Associates, Inc., 1966.

10. Fujisaki, et.al., 'Models for the word and sentence pitch contours of Japanese research on information processing', Annual Report No.2, Tokyo University, pp. 215-221.

11. Gilli L. et.al., 'Sequential system for recognizing spoken digits in real time', Acustica, 1967/68, 19 No.1.

12. Garding E, et.al., 'A study of the perception of some American English intonation contours', Paper read at 75th Annual meeting of the Modern Language Association of America, Philadelphia.

13. Gold B., 'Word recognition computer program', Res. Lab. for Electronics, MIT, Rept.No.452, June 1966.

14. Golden, R.M., 'Digital filter synthesis by sampled data transformation', IEEE Trans. Audio and Electro., Vol.AU-16, Sept. 1968, pp. 321-329.

15. Golden R.M., et.al., 'Design of wideband sampled data filters', BSTJ, Vol.43, July 1964, pp. 1533-1546.

16. Hadding-Koch et.al., 'An experimental study of some intonation contours', Phonatica, 1964, pp.175-185.

17. Hyde, S.R., 'Automatic speech recognition - literature survey and discussion', Research Dept. Report No.45, P.O. Research Dept., Dollis Hill, London.

18. Koing, W. etal. 'The sound spectrograph', JASA 18, July 1946, pp. 19-49.

19. Kuo and Kaiser, System Analysis by Digital Computer, John Wiley and Sons, 1965.

20. Liberman, P., Intonation, Perception and Language, Cambridge Mass, MIT Press.

21. Nagata, et.al., 'Spoken digit recognizer for Japanese language', NEC Research and Development, 1963, No.6.

22. Nelson, et.al, 'Acoustic recognition by analog feature abstraction techniques', Proc.of the Symposium on models for perception of speech and visual form, MIT Press,1964.

23. Newell, et.al, 'Speech understanding systems', Report Carnegie-Mellon University, Pittsberg, 1971.

24. Rabiner and Levitt, 'Analysis of fundamental frequency contours in speech,' JASA, Vol.49, No.2, Feb. 1971, pp. 569-582.

25. Rader and Gold, Digital Processing of Signals, McGraw-Hill Book Company, 1969.

26. Ramasubramanian, et.al, 'Segmentation of speech signals - a phonological approach', Tech. Report, No.113, TIFR Bombay, July 1973.

27. Reddy D.R., 'Computer recognition of connected speech', JASA, 42, 1967, pp. 329-347.

28. Sebasteyan, 'Automatic recognition of spoken numerals', JASA, 32, 1960, pp. 1516(a).

29. Sinha, V.P., 'Programs for linear digital filtering of time series', Tech. Report, Dept.of Electrical Engg., I.I.T. Kanpur, Nov. 1968.

30. Suzuki, J. et.al., 'Recognition of Japanse vowels - Preliminary to the recognition of speech', Journal of Radio Res.Lab.Tokyo, 1961, pp. 193-142.

31. Storer, J.E., Passive Network Synthesis, John Wiley and Sons, New York, 1957.

32. Wienberg, L. Network Analysis and Synthesis', McGraw-Hill Book Company, New York, 1962.